PCT/JP00/00423
27.01.00

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

REC'D 14 FEB 2000
WIPO PCT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

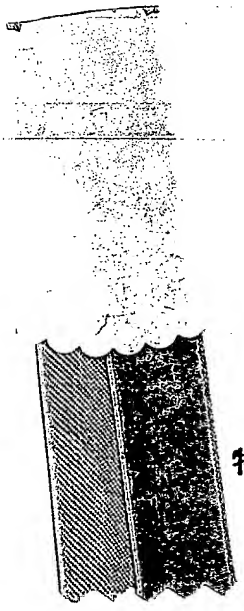
This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: １９９９年 １月２９日

出 願 番 号
Application Number: 平成１１年特許願第０２３０６４号

出 願 人
Applicant (s): ソニー株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)



１９９９年１１月１９日

特許庁長官
Commissioner,
Patent Office 近藤隆彦

出証番号 出証特平１１－３０８０１７４

【書類名】　特許願

【整理番号】　9801128002

【あて先】　特許庁長官殿

【国際特許分類】　G01B 11/24

【発明者】

【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内

【氏名】　トビー　ウォーカー

【発明者】

【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内

【氏名】　松原　弘

【特許出願人】

【識別番号】　000002185

【氏名又は名称】　ソニー株式会社

【代表者】　出井　伸之

【代理人】

【識別番号】　100067736

【弁理士】

【氏名又は名称】　小池　晃

【選任した代理人】

【識別番号】　100086335

【弁理士】

【氏名又は名称】　田村　榮一

【選任した代理人】

【識別番号】　100096677

【弁理士】

【氏名又は名称】　伊賀　誠司

【手数料の表示】

【予納台帳番号】　019530

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9707387

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　信号処理方法及び映像音声処理装置

【特許請求の範囲】

【請求項１】　供給された信号の内容の意味構造を反映するパターンを検出して解析する信号処理方法であって、

上記信号を構成する連続したフレームのひと続きから形成されるセグメントから、その特性を表す少なくとも１つ以上の特徴情報を抽出する特徴抽出工程と、

上記特徴情報を用いて、上記特徴情報のそれぞれ毎に、上記セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により上記セグメントの対の間の類似性を測定する類似性測定工程と、

上記特徴情報と上記測定基準とを用いて、上記セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つのセグメントを検出し、上記信号の内容の意味構造を反映し、時間的に連続するセグメントからなるシーンにまとめるグループ化工程とを備えること

を特徴とする信号処理方法。

【請求項２】　上記信号とは、ビデオデータにおける映像信号と音声信号との少なくとも１つであること

を特徴とする請求項１記載の信号処理方法。

【請求項３】　上記特徴抽出工程により、複数個に分割された単一のセグメント内の異なる時点における上記特徴情報の単一の統計的な代表値を選択して抽出すること

を特徴とする請求項１記載の信号処理方法。

【請求項４】　上記非類似性閾値は、複数個のセグメントの対の間の類似性の統計値を用いて決定されること

を特徴とする請求項１記載の信号処理方法。

【請求項５】　上記セグメントのうち、上記グループ化工程にてシーンにまとめられなかった少なくとも１つ以上のセグメントを、単一のシーンとしてまとめること

を特徴とする請求項１記載の信号処理方法。

【請求項６】　上記グループ化工程により得られた任意の特徴情報に関するシーンと、上記グループ化工程により得られた上記任意の特徴情報とは異なる特徴情報に関する少なくとも１つ以上のシーンとを単一にまとめること

を特徴とする請求項１記載の信号処理方法。

【請求項７】　上記グループ化工程により得られた上記映像信号における少なくとも１つ以上のシーンと、上記グループ化工程により得られた上記音声信号における少なくとも１つ以上のシーンとを単一にまとめること

を特徴とする請求項２記載の信号処理方法。

【請求項８】　供給されたビデオ信号の内容の意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置であって、

上記ビデオ信号を構成する連続した映像及び／又は音声フレームのひと続きから形成される映像及び／又は音声セグメントから、その特性を表す少なくとも１つ以上の特徴情報を抽出する特徴抽出手段と、

上記特徴情報を用いて、上記特徴情報のそれぞれ毎に、上記映像及び／又は音声セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により上記映像及び／又は音声セグメントの対の間の類似性を測定する類似性測定手段と、

上記特徴情報と上記測定基準とを用いて、上記映像及び／又は音声セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つの映像及び／又は音声セグメントを検出し、上記ビデオ信号の内容の意味構造を反映し、時間的に連続する映像及び／又は音声セグメントからなるシーンにまとめるグループ化手段とを備えること

を特徴とする映像音声処理装置。

【請求項９】　上記特徴抽出手段は、複数個に分割された単一の映像及び／又は音声セグメント内の異なる時点における上記特徴情報の単一の統計的な代表値を選択して抽出すること

を特徴とする請求項８記載の映像音声処理装置。

【請求項１０】　上記非類似性閾値は、複数個の映像及び／又は音声セグメントの対の間の類似性の統計値を用いて決定されること

を特徴とする請求項８記載の映像音声処理装置。

【請求項１１】　上記映像及び／又は音声セグメントのうち、上記グループ化手段によりシーンにまとめられなかった少なくとも１つ以上の映像及び／又は音声セグメントを、単一のシーンとしてまとめること

を特徴とする請求項８記載の映像音声処理装置。

【請求項１２】　上記グループ化手段により得られた任意の特徴情報に関するシーンと、上記グループ化手段により得られた上記任意の特徴情報とは異なる特徴情報に関する少なくとも１つ以上のシーンとを単一にまとめること

を特徴とする請求項８記載の映像音声処理装置。

【請求項１３】　上記グループ化工程により得られた上記ビデオ信号の映像信号における少なくとも１つ以上のシーンと、上記グループ化工程により得られた上記ビデオ信号の音声信号における少なくとも１つ以上のシーンとを単一にまとめること

を特徴とする請求項８記載の映像音声処理装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、信号の基礎となる意味構造を反映するパターンを検出して解析する信号処理方法及びビデオ信号の基礎となる意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置に関する。

【０００２】

【従来の技術】

例えばビデオデータに録画されたテレビ番組といった大量の異なる映像データにより構成される映像アプリケーションの中から、興味のある部分等の所望の部分を探して再生したい場合がある。

【０００３】

このように、所望の映像内容を抽出するための一般的な技術としては、アプリ

ケーションの主要場面を描いた一連の映像を並べて作成されたパネルであるストーリーボードがある。このストーリーボードは、ビデオデータをいわゆるショットに分解し、各ショットにおいて代表される映像を表示したものである。このような映像抽出技術は、そのほとんどが、例えば“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているように、ビデオストラクチャからショットを自動的に検出して抽出するものである。

【０００４】

【発明が解決しようとする課題】

ところで、例えば代表的な３０分のテレビ番組中には、数百ものショットが含まれている。そのため、上述した従来の映像抽出技術においては、ユーザが抽出された膨大な数のショットを並べたストーリーボードを調べる必要があり、このようなストーリーボードを理解するにはユーザに大きな負担を強いる必要があった。また、従来の映像抽出技術においては、例えば話し手の変化に応じて交互に２者を撮影した会話場面におけるショットは、冗長のものが多いという問題があった。このように、ショットは、ビデオストラクチャから抽出する対象としては階層が低すぎて必要のない情報量が多く、このようなショットを抽出する従来の映像抽出技術は、ユーザにとって利便性のよいものとはいえなかった。

【０００５】

また、他の映像抽出技術としては、例えば“A. Merlino, D. Morey and M. Maybury, Broadcast news navigation using story segmentation, Proc. of ACM Multimedia 97, 1997”や特開平１０－１３６２９７号公報に記載されているように、ニュースやフットボールゲームといった特定の内容ジャンルに関する非常に専門的な知識を用いるものがある。しかしながら、この従来の映像抽出技術は、目的のジャンルに関しては良好な結果を得ることができるが、他の種類のジャンルには全く役に立つものではなく、ジャンルに限定されて容易に一般化することができないという問題があった。

【０００６】

さらに、他の映像抽出技術としては、例えばU.S. Patent #5,708,767号公報に記載されているように、いわゆるストーリーユニットを抽出するものがある。しかしながら、この従来の映像抽出技術は、完全に自動化されたものではなく、どのショットが同じ内容を示すものであるかを決定するために、ユーザの介入が必要であった。また、この従来の映像抽出技術は、処理に要する計算が複雑であるとともに、適用対象として映像情報のみに限定されるといった問題もあった。

【０００７】

さらにまた、他の映像抽出技術としては、例えば特開平９－２１４８７９号公報に記載されているように、ショット検出と無音部分検出とを組み合わせることによりショットを識別するものがある。しかしながら、この従来の映像抽出技術は、２つの無音ポイント間にある１組のショットを識別するものであって、無音ポイントがショットの境界に対応する場合のみに限定されたものであった。

【０００８】

また、他の映像抽出技術としては、例えば“H. Aoki, S. Shimotsuji and O. Hori, A shot classification method to select effective key-frames for video browsing, IPSJ Human Interface SIG Notes, 7:43-50, 1996”や特開平９－９３５８８号公報に記載されているように、ストーリーボードにおける表示の冗長を低減するために反復されたショットを検出するものがある。しかしながら、この従来の映像抽出技術は、映像情報のみに適用できるものであり、音声情報に適用できるものではなかった。

【０００９】

本発明は、このような実情に鑑みてなされたものであり、上述した従来の映像抽出技術の問題を解決し、種々のビデオデータにおける高レベルのビデオストラクチャを抽出する信号処理方法及び映像音声処理装置を提供することを目的とするものである。

【００１０】

【課題を解決するための手段】

上述した目的を達成する本発明にかかる信号処理方法は、供給された信号の内

容の意味構造を反映するパターンを検出して解析する信号処理方法であって、信号を構成する連続したフレームのひと続きから形成されるセグメントから、その特性を表す少なくとも１つ以上の特徴情報を抽出する特徴抽出工程と、特徴情報を用いて、特徴情報のそれぞれ毎に、セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準によりセグメントの対の間の類似性を測定する類似性測定工程と、特徴情報と測定基準とを用いて、セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つのセグメントを検出し、信号の内容の意味構造を反映し、時間的に連続するセグメントからなるシーンにまとめるグループ化工程とを備えることを特徴としている。

【００１１】

このような本発明にかかる信号処理方法は、信号において類似したセグメントを検出してシーンにまとめる。

【００１２】

また、上述した目的を達成する本発明にかかる映像音声処理装置は、供給されたビデオ信号の内容の意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置であって、ビデオ信号を構成する連続した映像及び／又は音声フレームのひと続きから形成される映像及び／又は音声セグメントから、その特性を表す少なくとも１つ以上の特徴情報を抽出する特徴抽出手段と、特徴情報を用いて、特徴情報のそれぞれ毎に、映像及び／又は音声セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により映像及び／又は音声セグメントの対の間の類似性を測定する類似性測定手段と、特徴情報と測定基準とを用いて、映像及び／又は音声セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つの映像及び／又は音声セグメントを検出し、ビデオ信号の内容の意味構造を反映し、時間的に連続する映像及び／又は音声セグメントからなるシーンにまとめるグループ化手段とを備えることを特徴としている。

【００１３】

このように構成された本発明にかかる映像音声処理装置は、ビデオ信号におい

て類似した映像及び／又は音声セグメントを検出してまとめ、シーンとして出力する。

【００１４】

【発明の実施の形態】

以下、本発明を適用した具体的な実施の形態について図面を参照しながら詳細に説明する。

【００１５】

本発明を適用した実施の形態は、録画されたビデオデータから所望の内容を自動的に探し出して抽出する映像音声処理装置である。この映像音声処理装置の具体的な説明を行う前に、ここではまず本発明において対象とするビデオデータに関する説明を行う。

【００１６】

本発明において対象とするビデオデータについては、図１に示すようにモデル化し、フレーム、セグメント、シーンの３つのレベルに階層化されたストラクチャを有するものとする。すなわち、ビデオデータは、最下位層において、一連のフレームにより構成される。また、ビデオデータは、フレームの１つ上の階層として、連続するフレームのひと続きから形成されるセグメントにより構成される。さらに、ビデオデータは、最上位層において、このセグメントを意味のある関連に基づきまとめて形成されるシーンにより構成される。

【００１７】

このビデオデータは、映像及び音声の両方の情報を含むものである。すなわち、このビデオデータにおいてフレームは、単一の静止画像である映像フレームと、一般に数十～数百ミリセカンド／長といった短時間においてサンプルされた音声情報を表す音声フレームとを含むものである。

【００１８】

また、セグメントは、単一のカメラにより連続的に撮影された映像フレームのひと続きからなり一般にショットと呼ばれる映像セグメントと音声セグメントとを含むものである。このセグメントは、ビデオストラクチャにおける基本となる単位である。セグメントのうち、音声セグメントは、多くの定義が可能とされ、

例えば、以下に示すようなものが考えられる。例えば、音声セグメントは、一般によく知られている方法により検出されたビデオデータ中の無音期間により境界を定められて形成されるものがある。また、音声セグメントは、“D. Kimber and L. Wilcox, Acoustic Segmentation for Audio Browsers, Xerox Parc Technical Report”に記載されているように、例えば、音声、音楽、ボイド、無音等のように少数のカテゴリに分類された音声フレームのひと続きから形成されるものがある。さらに、音声セグメントは、“S. Pfeiffer, S. Fischer and E. Wolfgang, Automatic Audio Content Analysis, Proceeding of ACM Multimedia 96, Nov. 1996, pp21-30”に記載されているように、２枚の連続する音声フレーム間のある特性における大きな変化を検出する音声カット検出を用いて決定されるものがある。

【００１９】

さらに、シーンは、ビデオデータの内容を意味のあるより高いレベルでまとめてグループ化するために、映像セグメント（ショット）検出或いは音声セグメント検出により映像フレーム或いは音声フレームを捕捉したセグメントを、例えばセグメント内の知覚的アクティビティ量といったセグメントの特性を表す特徴情報であるフィーチャを用いて意味のあるまとまりにグループ化したものである。シーンは、主観的でありビデオデータの内容のジャンルに依存するものであるが、ここでは、映像又は音声における幾つかの関連性のあるフィーチャが互いに類似性を示す映像セグメント又は音声セグメントの反復パターンをグループ化したものとする。具体的には、図２に示すように、２人の話し手が互いに会話している場面で、映像セグメントは、話し手に応じて交互に現れる。このような反復パターンを有するビデオデータにおいて、一方の話し手における一連の映像セグメントＡと、他方の話し手における一連の映像セグメントＢとは、グループにまとめられて１つのシーンを構成する。このような反復パターンは、ビデオデータにおける高いレベルでの意味のあるストラクチャと非常に関係があり、シーンは、このようなビデオデータにおける高いレベルでの意味のあるまとまりを示すものである。

【００２０】

本発明を適用した実施の形態として図３に示す映像音声処理装置１０は、上述したビデオデータにおけるセグメントのフィーチャを用いてセグメント間の類似性を測定し、これらのセグメントをシーンにまとめてビデオストラクチャを自動的に抽出するものであり、映像セグメント及び音声セグメントの両方に適用できるものである。

【００２１】

映像音声処理装置１０は、同図に示すように、入力したビデオデータのストリームを映像、音声又はこれらの両方のセグメントに分割するビデオ分割部１１と、ビデオデータの分割情報を記憶するビデオセグメントメモリ１２と、各映像セグメントにおけるフィーチャを抽出する特徴抽出手段である映像フィーチャ抽出部１３と、各音声セグメントにおけるフィーチャを抽出する特徴抽出手段である音声フィーチャ抽出部１４と、映像セグメント及び音声セグメントのフィーチャを記憶するセグメントフィーチャメモリ１５と、映像セグメント及び音声セグメントをシーンにまとめるグループ化手段であるシーン検出部１６と、２つのセグメント間の類似性を測定する類似性測定手段であるフィーチャ類似性測定部１７とを備える。

【００２２】

ビデオ分割部１１は、例えば、ＭＰＥＧ１（Moving Picture Experts Group phase 1）やＭＰＥＧ２（Moving Picture Experts Group phase 2）、いわゆるＤＶ（Digital Video）のような圧縮ビデオデータフォーマットを含む種々のディジタル化されたフォーマットにおける映像データと音声データとからなるビデオデータのストリームを入力し、このビデオデータを映像、音声又はこれらの両方のセグメントに分割するものである。このビデオ分割部１１は、入力したビデオデータが圧縮フォーマットであった場合、この圧縮ビデオデータを完全伸張することなく直接処理することができる。ビデオ分割部１１は、入力したビデオデータを処理し、映像セグメントと音声セグメントとに分割する。また、ビデオ分割部１１は、入力したビデオデータを分割した結果である分割情報を後段のビデオセグメントメモリ１２に出力する。さらに、ビデオ分割部１１は、映像セグメン

トと音声セグメントとに応じて、分割情報を後段の映像フィーチャ抽出部１３及び音声フィーチャ抽出部１４に出力する。

【００２３】

ビデオセグメントメモリ１２は、ビデオ分割部１１から供給されたビデオデータの分割情報を記憶する。また、ビデオセグメントメモリ１２は、後述するシーン検出部１６からの問い合わせに応じて、分割情報をシーン検出部１６に出力する。

【００２４】

映像フィーチャ抽出部１３は、ビデオ分割部１１によりビデオデータを分割して得た各映像セグメント毎のフィーチャを抽出する。映像フィーチャ抽出部１３は、圧縮映像データを完全伸張することなく直接処理することができる。映像フィーチャ抽出部１３は、抽出した各映像セグメントのフィーチャを後段のセグメントフィーチャメモリ１５に出力する。

【００２５】

音声フィーチャ抽出部１４は、ビデオ分割部１１によりビデオデータを分割して得た各音声セグメント毎のフィーチャを抽出する。音声フィーチャ抽出部１４は、圧縮音声データを完全伸張することなく直接処理することができる。音声フィーチャ抽出部１４は、抽出した各音声セグメントのフィーチャを後段のセグメントフィーチャメモリ１５に出力する。

【００２６】

セグメントフィーチャメモリ１５は、映像フィーチャ抽出部１３及び音声フィーチャ抽出部１４からそれぞれ供給された映像セグメント及び音声セグメントのフィーチャを記憶する。セグメントフィーチャメモリ１５は、後述するフィーチャ類似性測定部１７からの問い合わせに応じて、記憶しているフィーチャの値やセグメントをフィーチャ類似性測定部１７に出力する。

【００２７】

シーン検出部１６は、ビデオセグメントメモリ１２に保持された分割情報と、１対のセグメント間の類似性とを用いて映像セグメント及び音声セグメントをそれぞれシーンにまとめる。シーン検出部１６は、グループ内の各セグメントから

開始して、セグメント群の中から類似しているセグメントの反復パターンを検出し、このようなセグメントを同一シーンとしてまとめてグループ化する。このシーン検出部１６は、あるシーンにおけるセグメントをまとめてグループを徐々に大きくしていき、全てのセグメントをグループ化するまで処理を行い、最終的に検出シーンを生成して出力する。シーン検出部１６は、フィーチャ類似性測定部１７を用いて、２つのセグメントがどの程度類似しているかを判断する。

【００２８】

フィーチャ類似性測定部１７は、２つのセグメント間の類似性を測定する。フィーチャ類似性測定部１７は、あるセグメントに関するフィーチャの値を検索するようにセグメントフィーチャメモリ１５に問いかける。

【００２９】

時間的に近接して反復している類似したセグメントは、ほぼ同一シーンの一部であるため、映像音声処理装置１０は、このようなセグメントを検出してグループ化していくことによって、シーンを検出する。このような映像音声処理装置１０は、図４に概略を示すような一連の処理を行うことによって、シーンを検出する。

【００３０】

まず、映像音声処理装置１０は、同図に示すように、ステップＳ１において、ビデオ分割を行う。すなわち、映像音声処理装置１０は、ビデオ分割部１１に入力されたビデオデータを映像セグメント又は音声セグメントのいずれか、或いは可能であればその両方に分割する。映像音声処理装置１０は、適用するビデオ分割方法に特に前提要件を設けない。例えば、映像音声処理装置１０は、“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているような方法によりビデオ分割を行う。このようなビデオ分割の方法は、当該技術分野ではよく知られたものであり、映像音声処理装置１０は、いかなるビデオ分割方法も適用できるものとする。

【００３１】

次に、映像音声処理装置１０は、ステップＳ２において、フィーチャの抽出を

行う。すなわち、映像音声処理装置１０は、映像フィーチャ抽出部１３や音声フィーチャ抽出部１４により各セグメントについて１組のフィーチャを計算し、セグメントの特性を表す。映像音声処理装置１０は、例えば、各セグメントの時間長や、カラーヒストグラムやテクスチャフィーチャといった映像フィーチャや、周波数解析結果、レベル、ピッチといった音声フィーチャやアクティビティ測定結果等を適用可能なフィーチャとして計算する。勿論、映像音声処理装置１０は、適用可能なフィーチャとしてこれらに限定されるものではない。

【００３２】

さらに、映像音声処理装置１０は、ステップＳ３において、フィーチャを用いたセグメントの類似性測定を行う。すなわち、映像音声処理装置１０は、フィーチャ類似性測定部１７により非類似性測定を行い、その測定基準により２つのセグメントがどの程度類似しているかを測定する。映像音声処理装置１０は、先のステップＳ２において抽出したフィーチャを用いて、非類似性測定基準を計算する。

【００３３】

そして、映像音声処理装置１０は、ステップＳ４において、セグメントのグループ化を行う。すなわち、映像音声処理装置１０は、先のステップＳ３において計算した非類似性測定基準と、先のステップＳ２において抽出したフィーチャとを用いて、時間的に近接して類似したセグメントを繰り返しまとめ、これらのセグメントを重ねてグループに合わせる。映像音声処理装置１０は、このようにして最終的に生成されたグループを検出シーンとして出力する。

【００３４】

このような一連の処理を経ることによって、映像音声処理装置１０は、ビデオデータからシーンを検出することができる。したがって、ユーザは、この結果を用いることによって、ビデオデータの内容を要約したり、ビデオデータ中の興味のあるポイントに迅速にアクセスしたりすることが可能となる。

【００３５】

以下、同図に示した映像音声処理装置１０における処理を各工程毎により詳細に説明していく。

【００３６】

まず、ステップＳ１におけるビデオ分割について説明する。映像音声処理装置１０は、ビデオ分割部１１に入力されたビデオデータを映像セグメント又は音声セグメントのいずれか、或いは可能であればその両方に分割するが、このビデオデータにおけるセグメントの境界を自動的に検出するための技術は、多くのものがあり、映像音声処理装置１０において、このビデオ分割方法に特別な前提要件を設けないことは上述した通りである。しかしながら、映像音声処理装置１０において、後の工程によるシーン検出の精度は、本質的に、基礎となるビデオ分割の精度に依存する。なお、映像音声処理装置１０におけるシーン検出は、ある程度ビデオ分割時のエラーを許容することができる。特に、映像音声処理装置１０において、ビデオ分割は、セグメント検出が不十分である部分よりも、セグメント検出が過度である部分においてエラーを生じる方が好ましい。映像音声処理装置１０は、類似したセグメントの検出が過度である結果である限り、たとえセグメント検出が過度でなかったとしても、一般に、シーン検出の際に検出過度であるセグメントを同一シーンとしてまとめることができる。

【００３７】

つぎに、ステップＳ２におけるフィーチャ抽出について説明する。フィーチャは、セグメントの内容を表すとともに、異なるセグメント間の類似性を測定するためのデータを供給するセグメントの属性である。映像音声処理装置１０は、映像フィーチャ抽出部１３や音声フィーチャ抽出部１４により各セグメントについて１組のフィーチャを計算し、セグメントの特性を表す。映像音声処理装置１０は、いかなるフィーチャの具体的詳細にも依存するものではない。映像音声処理装置１０において用いて効果的であると考えられるフィーチャとしては、例えば以下に示す映像フィーチャ、音声フィーチャ、映像音声共通フィーチャのようなものがある。映像音声処理装置１０において適用可能であるこのようなフィーチャは、非類似性測定を有していることが唯一の必要条件である。また、映像音声処理装置１０は、実際には、フィーチャ抽出と上述したビデオ分割とを効率化のために同時に行うことがあり、以下に説明するフィーチャは、このような処理を可能にするものである。

【００３８】

フィーチャとしては、まず映像フィーチャが挙げられる。セグメントを構成する映像（画像）は、そのセグメントの描写内容の大部分を表している。そのため、映像セグメントの類似性は、映像そのものの類似性に変えることができる場合が多い。したがって、映像フィーチャは、映像音声処理装置１０で用いることができる重要なフィーチャの１種である。この映像フィーチャは、動的な情報よりも静的な情報を表すため、映像音声処理装置１０は、後述するような方法によって、動的情報を得るように映像セグメント内の映像フィーチャを抽出する。

【００３９】

映像フィーチャとして既知のものは、多数存在するが、シーン検出のためには以下に示す色フィーチャ（ヒストグラム）及び映像相関が、計算コストと精度との良好な兼ね合いを与えることを見出したことから、映像音声処理装置１０は、映像フィーチャとしてこの色フィーチャ及び映像相関を用いる。

【００４０】

映像音声処理装置１０において、映像における色は、２つの映像が類似しているかを判断する際の重要な材料となる。カラーヒストグラムを用いて映像の類似性を判断することは、例えば“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”に記載されているように、よく知られている。カラーヒストグラムは、例えばＨＳＶやＲＧＢ等の３次元色空間をｎ個の領域に分割し、各領域にある映像における画素の相対的割合を計算したものである。そして、得られた情報からは、ｎ次元ベクトルが与えられる。圧縮されたビデオデータについては、例えばU.S. Patent #5,708,767号公報に記載されているように、カラーヒストグラムは、圧縮データから直接抽出される。

【００４１】

映像音声処理装置１０は、セグメントを構成する映像におけるもともとのＹＵＶ色空間を、色チャンネル当たり２ビットでサンプルし、長さ$2^{2\cdot3}$＝６４のヒストグラムを得る。

【００４２】

このようなヒストグラムは、映像の全体的な色調を表すが、空間的情報を欠いている。そこで、映像音声処理装置１０においては、映像相関を映像フィーチャの１つとして計算する。映像音声処理装置１０におけるシーン検出に関しては、インターリーブされたセグメントは、シーンストラクチャの有力な指標を与える。例えば会話場面において、カメラの位置は、２人の話し手の間を交互に移動するが、カメラは、通常、同一の話し手を再度撮影するときには同じ位置に戻る。このような場合を検出するためには、映像のグレイスケールの相関がセグメントの類似性の良好な指標となることを見出したことから、映像音声処理装置１０は、映像相関を計算するために、映像をＭ×Ｎの大きさのグレイスケール映像にサブサンプルする。ここで、ＭとＮは、両方とも小さい値であり、例えば８×８である。小さいグレイスケール映像は、長さＭＮのフィーチャとして解釈される。

【００４３】

さらに上述した映像フィーチャの他のフィーチャとしては、音声フィーチャが挙げられる。音声フィーチャは、音声セグメントの内容を表すのに用いることができるものである。映像音声処理装置１０は、音声フィーチャとして周波数解析、ピッチ、レベルを用いる。これらの音声フィーチャは、種々の文献により知られているものである。

【００４４】

まず、映像音声処理装置１０は、フーリエ変換等の周波数解析を行うことによって、単一の音声フレームにおける周波数情報の分布を決定することができる。映像音声処理装置１０は、例えば、１つの音声セグメントにわたる周波数情報の分布を表すために、ＦＦＴ（Fast Fourier Transform；高速フーリエ変換）成分、周波数ヒストグラム、パワースペクトル、その他のフィーチャを用いることができる。

【００４５】

また、映像音声処理装置１０は、平均ピッチや最大ピッチといったピッチや、平均ラウドネスや最大ラウドネスといったレベルを音声セグメントを表す有効な音声フィーチャとして用いることができる。

【００４６】

さらに他のフィーチャとしては、映像音声共通フィーチャが挙げられる。映像音声共通フィーチャは、特に映像フィーチャでもなく音声フィーチャでもないが、映像音声処理装置１０において、シーン内のセグメントの内容を表すのに有用な情報を与えるものである。映像音声処理装置１０は、この映像音声共通フィーチャとして、セグメント長とアクティビティとを用いる。

【００４７】

映像音声処理装置１０は、映像音声共通フィーチャとして、セグメント長を用いる。このセグメント長は、セグメントにおける時間長である。映像音声処理装置１０において、シーンは、よくその変化のリズムの特性を有し、シーン内のセグメント長の傾向と同様の傾向をともなう。例えば、迅速に連なった短いセグメントは、コマーシャルを表す。ところが、会話場面におけるセグメントは、より長くなるが、その長さの中でお互いに類似している。映像音声処理装置１０は、このような特性を有するセグメント長を映像音声共通フィーチャとして用いることができる。

【００４８】

また、映像音声処理装置１０は、映像音声共通フィーチャとして、アクティビティを用いる。アクティビティは、セグメントの内容がどの程度動的或いは静的であるように感じられるかを表す。例えば、アクティビティは、視覚的に動的である場合には、カメラが対象物に沿って迅速に移動する度合若しくは撮影されている物が迅速に変化する度合を表す。

【００４９】

このアクティビティは、カラーヒストグラムのようなフィーチャの平均フレーム間非類似性を測定することにより間接的に計算される。ここで、フレームｉとフレームｊとの間で測定されたフィーチャＦに対する非類似性測定基準をd_F（ｉ，ｊ）と定義すると、映像アクティビティV_Fは、次式（１）のように定義される。

【００５０】

【数１】

$$V_F = \frac{\sum_{i=b}^{f-1} d_F(i,i+1)}{f-b+1} \quad \cdots (1)$$

【００５１】

この式（１）において、ｂとｆは、それぞれ、１セグメントにおける最初と最後のフレームのフレーム番号である。映像音声処理装置１０は、具体的には、例えば上述したヒストグラムを用いて、映像アクティビティV_Fを計算する。

【００５２】

ところで、上述した映像フィーチャを始めとするフィーチャは、基本的にセグメントの静的情報を表すものであることは上述した通りであるが、セグメントの特徴を正確に表すためには、動的情報を表す必要がある。そこで、映像音声処理装置１０は、以下に示すようなフィーチャのサンプリング方法により動的情報を表す。

【００５３】

映像音声処理装置１０は、例えば図５に示すように、１セグメント内の異なる時点から１以上の静的なフィーチャ値の組を抽出する。このとき、映像音声処理装置１０は、フィーチャの抽出数を、忠実度の最大化と冗長度の最小化との釣り合いをとることにより決定する。例えば、セグメント内の１画像がキーフレームとして指定された場合には、そのキーフレームについて計算されたヒストグラムが抽出されたフィーチャとなる。

【００５４】

映像音声処理装置１０は、サンプリング方法を用いて、セグメントにおいて、あるフィーチャがとる全ての値のうち、どのサンプルを選択するかを決定する。映像音声処理装置１０は、優れたサンプリング方法を必要とする。

【００５５】

ここで、あるサンプルが常に所定の時点、例えばセグメント内の最後の時点に

おいてとられる場合を考える。この場合、黒にフェードする任意の２つのセグメントについては、サンプルが同一の黒いフレームとなるため、同一のフィーチャ値が得られる結果になる恐れがある。すなわち、これらのセグメントの映像内容がいかなるものであれ、サンプルした２つのフレームは、極めて類似していると判断されてしまう。このような問題は、サンプルが良好な代表値でないために発生するものである。

【００５６】

そこで、映像音声処理装置１０は、このように固定点でフィーチャを抽出するのではなく、統計的な代表値を抽出する。ここでは、一般的なフィーチャのサンプリング方法を２つの場合、すなわち、第１の場合として、フィーチャを実数のｎ次元ベクトルとして表すことができる場合と、第２の場合として、非類似性測定基準しか利用できない場合とについて説明する。なお、第１の場合は、ヒストグラムやパワースペクトル等、最もよく知られている映像フィーチャ及び音声フィーチャを含むものである。

【００５７】

第１の場合には、サンプル数は、事前にｋと決められており、映像音声処理装置１０は、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されてよく知られているｋ平均値クラスタリング法（k-means-clustering method）を用いて、セグメント全体についてのフィーチャ値をｋ個の異なるグループに自動的に分割する。そして、映像音声処理装置１０は、ｋ個の各グループから１サンプルを選択する。すなわち、映像音声処理装置１０は、グループのセントロイド（平均ベクトル）又はグループのセントロイドに近いサンプルを選択する。映像音声処理装置１０は、この処理を短時間で行うことができ、サンプル数において直線的な時間しか要さない。

【００５８】

一方、第２の場合には、映像音声処理装置１０は、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されているｋ－メドイドアルゴリズム法（k-me

doids algorithm method）を用いて、ｋ個のグループを形成する。そして、映像音声処理装置１０は、各グループ毎に、サンプル値として、上述したグループのセントロイドに類似するグループのメドイドを用いる。

【００５９】

なお、映像音声処理装置１０においては、抽出されたフィーチャについての非類似性測定基準を構成する方法は、その基礎となるフィーチャの非類似性測定基準に基づくが、これについては後述する。

【００６０】

このようにして、映像音声処理装置１０は、静的なフィーチャを抽出し、動的情報を表すことができる。

【００６１】

以上のように、映像音声処理装置１０は、種々のフィーチャを抽出することができる。これらの各フィーチャは、一般に、単一ではセグメントの内容を表すのに不十分であることが多い。そこで、映像音声処理装置１０は、これらの各種フィーチャを組み合わせることで、互いに補完し合うフィーチャの組を選択することができる。例えば、映像音声処理装置１０は、上述したカラーヒストグラムと映像相関とを組み合わせることによって、各フィーチャが有する情報よりも多くの情報を得ることができる。

【００６２】

つぎに、図４中ステップＳ３におけるフィーチャを用いたセグメントの類似性測定について説明する。映像音声処理装置１０は、２つのフィーチャについて、その２つのフィーチャ値がどの程度非類似であるかを測定する実数値を計算する関数である非類似性測定基準を用いて、フィーチャ類似性測定部１７によりセグメントの類似性測定を行う。この非類似性測定基準は、その値が小さい場合は２つのフィーチャが類似していることを示し、値が大きい場合は非類似であることを示す。ここでは、フィーチャＦに関する２つのセグメントS_1，S_2の非類似性を計算する関数を非類似性測定基準d_F（S_1，S_2）と定義する。このような関数は、以下の式（２）に示す特性に合致するものでなければならない。

【００６３】

【数２】

$$d_F(S_1,S_2)=0 \qquad (S_1=S_2 \text{のとき})$$
$$d_F(S_1,S_2)\geq 0 \qquad (\text{全ての}S_1,S_2\text{について}) \qquad \cdots(2)$$
$$d_F(S_1,S_2)=d_F(S_2,S_1) \qquad (\text{全ての}S_1,S_2\text{について})$$

【００６４】

なお、適切な非類似性測定基準が、特定のフィーチャに依存することがあるが、“G. Ahanger and T.D.C. Little, A survey of technologies for parsing and indexing digital video, J. of Visual Communication and Image Representation 7:28-4, 1996”や“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis, John-Wiley and sons, 1990”に記載されているように、多くの一般的な非類似性測定基準は、ｎ次元空間における点として表されるフィーチャについての類似性を測定するのに有効であるとして知られている。これらは、ユークリッド距離、内積、Ｌ１距離等である。特に、Ｌ１距離が、ヒストグラムや映像相関といったフィーチャを含む種々のフィーチャに関して、有効に作用することを見出したことから、映像音声処理装置１０は、２つのｎ次元ベクトルＡ，Ｂ間のＬ１距離を次式（３）のように求める。

【００６５】

【数３】

$$d_{L1}=(A,B)=\sum_{i=1}^{n}\left|A_i-B_i\right| \qquad \cdots(3)$$

【００６６】

ここで、下付文字は、ｎ次元ベクトルのｉ番目の要素を示すものである。

【００６７】

また、映像音声処理装置１０は、上述したように、時間とともに変化するフィ

ーチャに関して、セグメントにおける様々な時点でフィーチャ値を抽出する。そして、映像音声処理装置１０は、２つの抽出されたフィーチャの間の類似性を決定するために、抽出されたフィーチャの非類似性を、その基礎となるフィーチャの非類似性測定基準により定義する。多くの場合、非類似性は、２つの抽出されたフィーチャのそれぞれから選択された最も非類似性のないフィーチャの対の非類似性値を用いて決定されるのが最良である。すなわち、映像音声処理装置１０は、最小非類似性を決定する。この場合、２つの抽出されたフィーチャSF_1, SF_2の間のサンプルされた非類似性測定基準は、次式（４）のように定義される。

【００６８】

【数４】

$$d_s(SF_1, SF_2) = \min_{F_1 \in SF_1, F_2 \in SF_2} d_F(F_1, F_2) \qquad \cdots (4)$$

【００６９】

上式（４）における関数d_F（F_1, F_2）は、その基礎となる抽出されたフィーチャＦについての非類似性測定基準を示す。また、非類似性の値の最小をとる代わりに、最大又は平均をとる方がより適切である場合もある。

【００７０】

さらに、映像音声処理装置１０は、多くの場合、類似性を決定するために、同一セグメントについての多数のフィーチャから情報を組み合わせる必要がある。この１つの方法として、映像音声処理装置１０は、種々のフィーチャベクトル非類似性関数の重み付きの組み合わせを計算する。すなわち、映像音声処理装置１０は、ｋ個のフィーチャF_1, F_2, ・・・, F_kが存在する場合、次式（５）に表される組み合わせたフィーチャに関する非類似性測定基準d_F（S_1, S_2）を用いる。

【００７１】

【数５】

$$d_F(S_1,S_2) = \Sigma_{i-1}^{k} w_i d_{Fi}(S_1,S_2) \qquad \cdots (5)$$

【００７２】

ここで、｛w_i｝は、重みと$\Sigma_i w_i = 1$との組である。

【００７３】

以上のように、映像音声処理装置１０は、図４中ステップＳ２において抽出したフィーチャを用いて非類似性測定基準を計算し、セグメントの類似性を測定することができる。

【００７４】

つぎに、図４中ステップＳ４におけるセグメントのグループ化について説明する。映像音声処理装置１０は、非類似性測定基準と抽出したフィーチャとを用いて、時間的に近接して類似したセグメントを繰り返しまとめ、これらのセグメントを重ねてグループに合わせ、最終的に生成されたグループを検出シーンとして出力する。

【００７５】

映像音声処理装置１０は、セグメントをグループ化してシーンを検出する際に、２つの基本的な処理を行う。映像音声処理装置１０は、まず第１の処理として、互いに時間的に近接して類似したセグメントのグループを検出する。この処理により得られるグループは、ほとんどが同一シーンの一部となるものである。また、映像音声処理装置１０は、セグメントが時間的に重複しているため、第２の処理として、互いに時間が重複して類似したシーンをまとめる。映像音声処理装置１０は、このような処理を各セグメントから開始し、反復して繰り返す。そして、映像音声処理装置１０は、徐々にセグメントのグループを大きく構築していき、最終的に生成したグループをシーンの組として出力する。

【００７６】

このような処理において、映像音声処理装置１０は、その処理動作を制御するために２つの制約を用いる。

【００７７】

すなわち、映像音声処理装置１０は、第１の制約として、２つのセグメントが、どの程度類似している場合に同一のシーンのものであるのに十分類似しているかと考えられるかを決定する非類似性閾値δ_{sim}を用いる。例えば、図６に示すように、映像音声処理装置１０は、あるセグメントに対して一方のセグメントが類似性領域に属するか非類似性領域に属するかを判断する。

【００７８】

なお、映像音声処理装置１０は、非類似性閾値δ_{sim}をユーザにより設定するようにしてもよく、また、後述するように、自動的に決定してもよい。

【００７９】

また、映像音声処理装置１０は、第２の制約として、２つのセグメントが、どの程度広く分割されている場合に同一のシーンのものであるかとなお考えられ得るかを決定する時間閾値Ｔを用いる。例えば、図７に示すように、映像音声処理装置１０は、時間閾値Ｔの範囲内で互いに近接して続いている類似した２つのセグメントＡ，Ｂを同一シーンにまとめるが、時間的に大きく離れていて時間閾値Ｔの範囲外である２つのセグメントＢ，Ｃをまとめることはない。このように、映像音声処理装置１０は、この時間閾値Ｔによる時間制約があるために、互いに類似しているものの大きく離れているセグメントを同一シーンにまとめてしまうエラーを発生することがない。

【００８０】

なお、この時間閾値Ｔとしては、６～８の場合が概して良好な結果を与えることを見出したことから、映像音声処理装置１０は、基本的に、時間閾値Ｔを６～８として用いる。

【００８１】

映像音声処理装置１０は、類似セグメントのグループを求めるために、ここでは、“L. Kaufman and P.J. Rousseeuw, Finding Groups in Data:An Introduct

ion to Cluster Analysis, John-Wiley and sons, 1990”に記載されている階層的クラスタ分割方法（hierarchical clustering method）を適合させて用いることにする。このアルゴリズムにおいては、２つのクラスタC_1，C_2間の非類似性測定基準$d_C(C_1, C_2)$について、次式（６）に示すように、含まれる２つの要素間の最小非類似性として定義する。

【００８２】

【数６】

$$d_C(C_1,C_2) \equiv \min_{S_1 \in C_1, S_2 \in C_2} dist_S(S_1,S_2) \qquad \cdots (6)$$

【００８３】

なお、映像音声処理装置１０においては、上式（６）で示される最小関数を最大又は平均関数に容易に置換することができる。

【００８４】

まず、映像音声処理装置１０は、図８に示すように、ステップＳ１１において、変数Ｎをセグメントの数に初期化する。この変数Ｎは、常に現在検出されているグループの数を示すものである。

【００８５】

次に、映像音声処理装置１０は、ステップＳ１２において、クラスタの組を生成する。映像音声処理装置１０は、初期時には、Ｎ個の各セグメントを異なるクラスタとする。そのため、初期時には、Ｎ個のクラスタが存在することになる。各クラスタは、C^{start}とC^{end}とによって表されるその開始時と終了時とを示す特性を有する。クラスタの組の要素は、C^{start}から順番に整頓される。

【００８６】

さらに、映像音声処理装置１０は、ステップＳ１３において、変数ｔを１に初期化し、ステップＳ１４において、変数ｔが時間閾値Ｔよりも大きいか否かを判別する。ここで、映像音声処理装置１０は、変数ｔが時間閾値Ｔよりも大きい場合には、ステップＳ２３へと処理を移行し、変数ｔが時間閾値Ｔよりも小さい場

合には、ステップＳ１５へと処理を移行する。ただし、ここでは、変数ｔが１であるため、映像音声処理装置１０は、ステップＳ１５へと処理を移行する。

【００８７】

映像音声処理装置１０は、ステップＳ１５において、非類似性測定基準d_Cを計算し、Ｎ個のクラスタの中から最も類似した２つのクラスタを検出する。ただし、ここでは、変数ｔが１であるため、映像音声処理装置１０は、隣接したクラスタ間の非類似性測定基準d_Cを計算し、その中から最も類似したクラスタの対を検出する。

【００８８】

このような最も類似した２つのクラスタを検出する方法としては、対象となる全てのクラスタの対を走査することが考えられる。映像音声処理装置１０は、時間閾値Ｔによる時間制約を有するため、対象とするクラスタの対の数を限定することができる。そのため、映像音声処理装置１０は、ｔ個のセグメントにより分割されているクラスタのみを走査すればよい。クラスタの組は、時間順に整頓されているため、映像音声処理装置１０は、あるクラスタに対してその前方及び後方へと走査していき、ｔ個のセグメントよりも離れたクラスタを走査すれば、このクラスタ以降の全てのセグメントを対象外として走査処理を終了することができる。

【００８９】

ここで、検出された２つのクラスタをそれぞれC_i，C_jと定義し、これらのクラスタC_i，C_jの間の非類似性の値をd_{ij}と定義する。

【００９０】

映像音声処理装置１０は、ステップＳ１６において、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも大きいか否かを判別する。ここで、映像音声処理装置１０は、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも大きい場合には、ステップＳ２１へと処理を移行し、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも小さい場合には、ステップＳ１７へと処理を移行する。ここでは、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも小さいものとする。

【００９１】

映像音声処理装置１０は、ステップＳ１７において、クラスタC_jをクラスタC_iに合わせる。すなわち、映像音声処理装置１０は、クラスタC_jの要素の全てをクラスタC_iに加える。

【００９２】

また、映像音声処理装置１０は、ステップＳ１８において、クラスタC_jをクラスタの組から除去する。なお、２つのクラスタC_i，C_jを合わせることにより開始時C_i^{start}の値が変化した場合には、映像音声処理装置１０は、開始時による順序を維持するために、クラスタの組の要素を再び並べ替える。

【００９３】

さらに、映像音声処理装置１０は、ステップＳ１９において、変数Ｎから１を減じる。

【００９４】

そして、映像音声処理装置１０は、ステップＳ２０において、変数Ｎが１であるか否かを判別する。ここで、映像音声処理装置１０は、変数Ｎが１である場合には、ステップＳ２３へと処理を移行し、変数Ｎが１でない場合には、ステップＳ１５へと処理を移行する。ここでは、変数Ｎが１でないものとする。

【００９５】

すると、映像音声処理装置１０は、ステップＳ１５において、再び非類似性測定基準d_Cを計算し、Ｎ－１個のクラスタの中から最も類似した２つのクラスタを検出する。ここでも、変数ｔが１であるため、映像音声処理装置１０は、隣接したクラスタ間の非類似性測定基準d_Cを計算し、その中から最も類似したクラスタの対を検出する。

【００９６】

また、映像音声処理装置１０は、ステップＳ１６において、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも大きいか否かを判別する。ここでも、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも小さいものとする。

【００９７】

そして、映像音声処理装置１０は、ステップＳ１７乃至ステップＳ２０の処理

を行う。

【００９８】

映像音声処理装置１０は、このような処理を繰り返し、変数Ｎが減算されていった結果、ステップＳ２０において、変数Ｎが１であると判別した場合には、ステップＳ２３において、単一のセグメントのみを含むクラスタを合わせる。すなわち、映像音声処理装置１０は、この場合は、全てのセグメントが１つのクラスタにまとめられているため、この処理を行う必要はなく、一連の処理を終了する。

【００９９】

さて、映像音声処理装置１０は、ステップＳ１６において、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも大きいと判別した場合には、ステップＳ２１へと処理を移行するが、この場合には、ステップＳ２１において、時間的に重複しているクラスタを繰り返し合わせる。すなわち、C_iの時間間隔$[C_i^{start}, C_i^{end}]$が、C_jの時間間隔$[C_j^{start}, C_j^{end}]$と相交している場合には、２つのクラスタC_iとC_jは、重複しているため、映像音声処理装置１０は、クラスタをその組の開始時により整頓すると、重複しているクラスタを検出し、１つに合わせることができる。

【０１００】

そして、映像音声処理装置１０は、ステップＳ２２において、変数ｔに１を加算してｔ＝２とし、ステップＳ１４へと処理を移行して変数ｔが時間閾値Ｔよりも大きいか否かを判別する。ここでも、変数ｔが時間閾値Ｔよりも小さいものとし、映像音声処理装置１０は、ステップＳ１５へと処理を移行するものとする。

【０１０１】

映像音声処理装置１０は、ステップＳ１５において、非類似性測定基準d_Cを計算し、現在存在する複数のクラスタの中から最も類似した２つのクラスタを検出する。ただし、ここでは、変数ｔが２であるため、映像音声処理装置１０は、１つおきに離れているクラスタ間の非類似性測定基準d_Cを計算し、その中から最も類似したクラスタの対を検出する。

【０１０２】

そして、映像音声処理装置１０は、ステップＳ１６において、１つおきに離れているクラスタC_i，C_jの非類似性値d_{ij}が非類似性閾値δ_{sim}よりも大きいか否かを判別する。ここでも、非類似性値d_{ij}が非類似性閾値δ_{sim}よりも大きいものとし、映像音声処理装置１０は、ステップＳ２１及びステップＳ２２の処理を経て、変数ｔに１を加算してｔ＝３として再びステップＳ１４以降の処理へと移行する。勿論、映像音声処理装置１０は、変数ｔが３のときには、ステップＳ１５において、２つおきに離れているクラスタ間の非類似性測定基準d_Cを計算し、その中から最も類似したクラスタの対を検出する。

【０１０３】

映像音声処理装置１０は、このような処理を繰り返し、変数ｔが加算されていった結果、ステップＳ１４において、変数ｔが時間閾値Ｔよりも大きいと判別すると、ステップＳ２３へと処理を移行し、単一のセグメントのみを含むクラスタを合わせる。すなわち、映像音声処理装置１０は、孤立しているクラスタを単一の要素のみを含むクラスタとみなし、このような一連のクラスタが存在している場合には、これらのクラスタを単一のクラスタに合わせる。この工程は、単一のシーンと類似性関連を有さないセグメントをまとめるものである。なお、映像音声処理装置１０は、必ずしもこの工程を行う必要はない。

【０１０４】

そして、映像音声処理装置１０は、ステップＳ２２の処理を経て一連の処理を終了する。

【０１０５】

このような一連の処理によって、映像音声処理装置１０は、複数のクラスタをまとめていき、検出シーンを生成することができる。

【０１０６】

なお、映像音声処理装置１０は、非類似性閾値δ_{sim}をユーザにより設定するようにしてもよく、自動的に決定してもよいことは上述した通りである。非類似性閾値δ_{sim}は、固定された値を用いる場合には、その最適な値がビデオデータの内容に依存する。非類似性閾値δ_{sim}は、例えば、変化に富んだ映像内容を有

する番組の場合には、高い値に設定される必要があり、変化が少ない映像内容を有する番組の場合には、より低い値に設定される必要がある。そして、非類似性閾値δ_{sim}が高すぎる場合には、検出されるシーンが少なくなりすぎ、非類似性閾値δ_{sim}が低すぎる場合には、検出されるシーンが多くなりすぎることになる。

【０１０７】

このようなことから、映像音声処理装置１０は、最適な非類似性閾値δ_{sim}を決定することが、その性能を左右する上で重要となる。そのため、映像音声処理装置１０は、非類似性閾値δ_{sim}をユーザにより設定する場合には、上述したことを考慮した上で設定する必要がある。また、映像音声処理装置１０は、以下に示す方法により、有効な非類似性閾値δ_{sim}を自動的に決定することもできる。

【０１０８】

すなわち、映像音声処理装置１０は、１つの方法として、平均値や最頻値といった統計的なフィーチャを用いて、（ｎ）（ｎ－１）／２個のセグメント対間の類似性の違いを表す距離の分布から非類似性閾値δ_{sim}を得る。例えば、映像音声処理装置１０は、全てのセグメント対における類似性の違いを表す距離の平均値μとその標準偏差σとを用いて、非類似性閾値δ_{sim}を$a\mu+b\sigma$と設定する。ここで、ａ及びｂは、ある固定定数であるが、この定数ａ及びｂは、それぞれ、０．５及び０．１に設定することが良好な結果を与えることを見出している。

【０１０９】

実際には、映像音声処理装置１０は、全てのセグメント対について、それらの間の類似性の違いを表す距離を決定する必要はない。映像音声処理装置１０は、全てのセグメント対を対象とする代わりに、平均値μ及び標準偏差σの真値を判断するための距離値の部分集合をランダムに抽出する。すなわち、映像音声処理装置１０は、距離を計算する際に、２つのセグメントをランダムに選択する。映像音声処理装置１０は、２つのセグメントを十分に選択した場合には、平均値μ及び標準偏差σが真値に非常に近い結果を得ることができ、適切な非類似性閾値δ_{sim}を自動的に得ることができる。このような場合、映像音声処理装置１０は、例えば、セグメントの全数をｎ、任意の小さい定数をＣとして、Ｃｎと表され

るような数のセグメント対を抽出することによって、適切な非類似性閾値δ_{sim}を自動的に決定することができる。

【０１１０】

また、映像音声処理装置１０は、シーンを検出する際に、セグメントが同一グループであるかを決定するために、単一の非類似性測定基準をもちいるばかりではなく、重み付け関数を用いて、異種のフィーチャに関する多様な非類似性測定基準を組み合わせることができることは上述した通りである。映像音声処理装置１０において、このようなフィーチャの重み付けは、試行錯誤の末得られるものであり、各フィーチャが質的に異なるタイプのものである場合には、通常、適切な重み付けを行うことは困難であるが、映像音声処理装置１０は、例えば、カラーヒストグラムとテクスチャフィーチャとを組み合わせる場合、各フィーチャに関してシーンを検出し、検出された各シーンストラクチャを単一のシーンストラクチャに組み合わせることができる。ここで、各フィーチャに関してシーンを検出した結果をシーン層と称することにする。例えば、フィーチャとしてカラーヒストグラムとセグメント長とを用いる場合、映像音声処理装置１０は、これらのフィーチャに対してシーンを検出し、カラーヒストグラムについてのシーン層と、セグメント長についてのシーン層とを得て、これらのシーン層を単一のシーンストラクチャに組み合わせることができる。

【０１１１】

さらに、一般に、映像領域と音声領域とからの情報を組み合わせることはできないが、映像音声処理装置１０は、質的に異なるタイプのフィーチャを組み合わせる場合と同様な方法により、映像領域と音声領域とから得られるシーン層を単一のシーンストラクチャに組み合わせることができる。

【０１１２】

このような処理について説明する。ここでは、それぞれが類似性の１つの基準を表すｋ個のフィーチャF_1, F_2, ・・・, F_kがあるものとし、各フィーチャF_iに対応して、非類似性測定基準$d_F{}^i$と、非類似性閾値δ^i_{sim}と、時間閾値T^iとがあるものとする。映像音声処理装置１０は、これらの各フィーチャF_iに対する非類似性測定基準$d_F{}^i$と、非類似性閾値δ^i_{sim}と、時間閾値T^iとを用いて

シーン層の組$X_i = \{X_i^j\}$を検出する。また、映像音声処理装置１０は、映像情報と音声情報とに対して分割的にシーン層を検出し、映像情報と音声情報とに関する２つの独立したシーン層$X_i = \{X_i^j\}$（ｉ＝１，２）を生成するものとする。

【０１１３】

映像音声処理装置１０は、異なるシーン層を単一のシーンストラクチャに組み合わせるため、シーン境界の組み合わせ方を決定する必要がある。このシーン境界は、お互いにそろっている保証はないものである。ここで、各シーン層に関して、シーン境界を示す一連の時間で表される境界点t_{i1}，t_{i2}，・・・，$t_{i|X_i|}$があるものとする。まず、映像音声処理装置１０は、種々のシーン層を単一のグループに組み合わせるために、最初にあるシーン層を整列に関する基礎とするために選択する。そして、映像音声処理装置１０は、最終的に組み合わせて生成するシーンストラクチャにおけるシーン境界かどうかを各境界点t_{i1}，t_{i2}，・・・，$t_{i|X_i|}$に関して決定する。

【０１１４】

ここで、シーン層の組の番号である各ｉに関して、$B_i(t)$を、ｉ番目のシーン層X_iにおいて、ある時間ｔで「近く」にシーン境界があるかどうかを示すブール関数とする。この「近く」の意味は、変化し、例えば、映像情報と音声情報とをそろえている場合には、０．５秒であったりする。

【０１１５】

映像音声処理装置１０は、各境界点$t_j = t_{ij}$，$j = 1, \cdots, |X_j|$に関して、$l = 1, \cdots, k$のそれぞれについて、関数$B_l(t_j)$の結果を計算する。この結果は、各独立したシーン層に関して、シーン層X_iにおいて時間ｔの近くにシーン境界があるかどうかを示している。そして、映像音声処理装置１０は、決定関数として、組み合わされた決定において時間ｔがシーン境界であるかどうかを決定するための$B_i(t_j)$の値を用いる。

【０１１６】

このような決定関数の１つの単純な例は、$B_i(t_j)$が定数ｍ以上の１に等しい値の総数を計算することであり、時間の点ｔは、最後のシーンストラクチャに

おけるシーン境界を明示する。特に、ｍ＝１の場合、これは、「ｏｒ－ｉｎｇ」境界点と同義のものである。ｍ＝ｋの場合、全ての境界点をそろえる要求と同義である。

【０１１７】

このようにして、映像音声処理装置１０は、異なるシーン層を単一のシーンストラクチャに組み合わせることができる。

【０１１８】

以上説明してきたように、本発明の実施の形態として示す映像音声処理装置１０は、シーンストラクチャを抽出するものであって、異なるタイプのビデオデータの内容に適用できるものである。映像音声処理装置１０は、実際のビデオデータの内容において試験されてきており、テレビドラマや映画といった他のビデオデータジャンルからも同様に、シーンストラクチャを復活可能であることが論証済みである。

【０１１９】

また、映像音声処理装置１０は、完全に自動的であり、上述した非類似性閾値や時間閾値を設定するために、ユーザの介入を必要とせず、ビデオデータの内容の変化に応じて、適切な閾値を自動的に決定することができる。

【０１２０】

さらに、映像音声処理装置１０は、ユーザが事前にビデオデータの内容のストラクチャを知る必要はない。

【０１２１】

さらにまた、映像音声処理装置１０は、非常に単純であり計算上の効率もよいため、セットトップボックスやディジタルビデオレコーダ、ホームサーバ等の家庭用電子機器にも適用することができる。

【０１２２】

また、映像音声処理装置１０は、シーンを検出した結果、ビデオブラウジングのための新たな高レベルアクセスの基礎を得ることができる。そのため、映像音声処理装置１０は、セグメントではなくシーンといった高レベルのビデオストラクチャを用いてビデオデータの内容を映像化することにより、内容に基づいたビ

デオデータへのアクセスが可能になる。例えば、映像音声処理装置１０は、シーンを表示することにより、ユーザは、番組の要旨をすばやく知ることができ、興味のある部分を迅速に見つけることができる。

【０１２３】

さらに、映像音声処理装置１０は、シーン検出の結果、ビデオデータの概要又はダイジェストを自動的に作成するための基盤が得られる。一般に、ビデオデータからのランダムな断片を組み合わせるのとは異なり、一貫した概要を作成するには、ビデオデータを、再構成可能な意味を持つ成分に分解することができることが必要である。映像音声処理装置１０により検出されたシーンは、そのようなダイジェストを作成するための正常な基盤である。

【０１２４】

なお、本発明は、上述した実施の形態に限定されるものではなく、例えば、セグメント間の類似性測定のために用いるフィーチャ等は、上述したもの以外でもよいことは勿論であり、その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。

【０１２５】

【発明の効果】

以上詳細に説明したように、本発明にかかる信号処理方法は、供給された信号の内容の意味構造を反映するパターンを検出して解析する信号処理方法であって、信号を構成する連続したフレームのひと続きから形成されるセグメントから、その特性を表す少なくとも１つ以上の特徴情報を抽出する特徴抽出工程と、特徴情報を用いて、特徴情報のそれぞれ毎に、セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準によりセグメントの対の間の類似性を測定する類似性測定工程と、特徴情報と測定基準とを用いて、セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つのセグメントを検出し、信号の内容の意味構造を反映し、時間的に連続するセグメントからなるシーンにまとめるグループ化工程とを備える。

【０１２６】

したがって、本発明にかかる信号処理方法は、信号において類似したセグメントを検出してシーンにまとめることができ、セグメントよりも高いレベルのストラクチャを抽出することができる。

【０１２７】

また、本発明にかかる映像音声処理装置は、供給されたビデオ信号の内容の意味構造を反映する映像及び／又は音声のパターンを検出して解析する映像音声処理装置であって、ビデオ信号を構成する連続した映像及び／又は音声フレームのひと続きから形成される映像及び／又は音声セグメントから、その特性を表す少なくとも１つ以上の特徴情報を抽出する特徴抽出手段と、特徴情報を用いて、特徴情報のそれぞれ毎に、映像及び／又は音声セグメントの対の間の類似性を測定する測定基準を算出して、この測定基準により映像及び／又は音声セグメントの対の間の類似性を測定する類似性測定手段と、特徴情報と測定基準とを用いて、映像及び／又は音声セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つの映像及び／又は音声セグメントを検出し、ビデオ信号の内容の意味構造を反映し、時間的に連続する映像及び／又は音声セグメントからなるシーンにまとめるグループ化手段とを備える。

【０１２８】

したがって、本発明にかかる映像音声処理装置は、ビデオ信号において類似した映像及び／又は音声セグメントを検出してまとめ、シーンとして出力することが可能であり、映像及び／又は音声セグメントよりも高いレベルのビデオストラクチャを抽出することが可能となる。

【図面の簡単な説明】

【図１】

本発明において適用するビデオデータの構成を説明する図であって、モデル化したビデオデータのストラクチャを説明する図である。

【図２】

シーンを説明する図である。

【図３】

本発明の実施の形態として示す映像音声処理装置の構成を説明するブロック図である。

【図４】

同映像音声処理装置において、シーンを検出してグループ化する際の一連の工程を説明するフローチャートである。

【図５】

同映像音声処理装置における動的フィーチャサンプリング処理を説明する図である。

【図６】

非類似性閾値を説明する図である。

【図７】

時間閾値を説明する図である。

【図８】

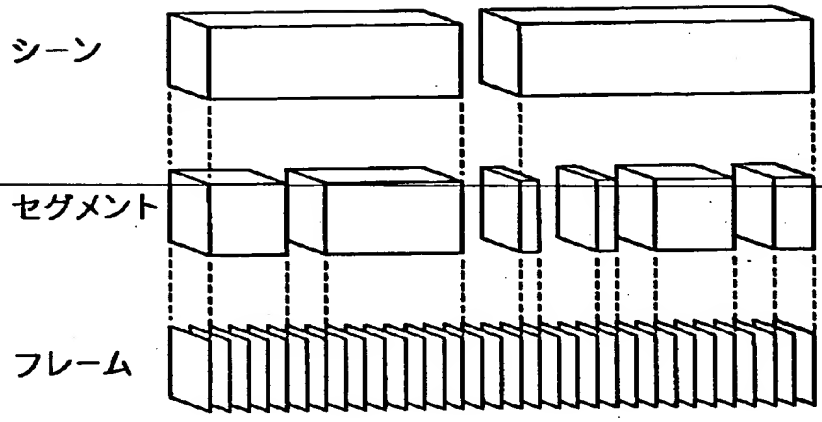
同映像音声処理装置において、セグメントをグループ化する際の一連の工程を説明するフローチャートである。

【符号の説明】

１０　映像音声処理装置、　１１　ビデオ分割部、　１２　ビデオセグメントメモリ、　１３　映像フィーチャ抽出部、　１４　音声フィーチャ抽出部、　１５　セグメントフィーチャメモリ、　１６　シーン検出部、　１７　フィーチャ類似性測定部

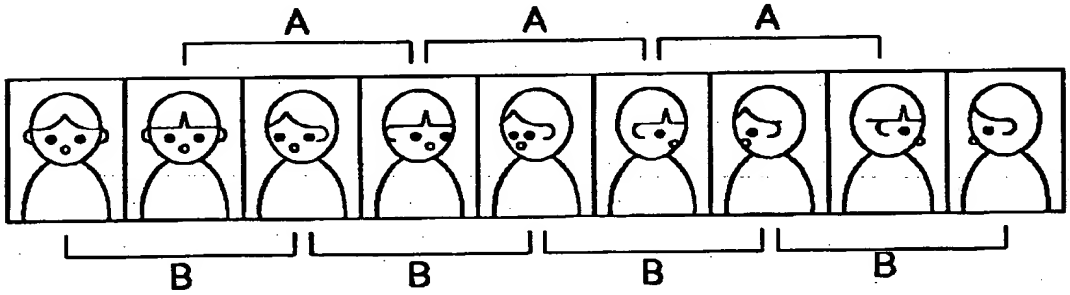
【書類名】　図面

【図１】



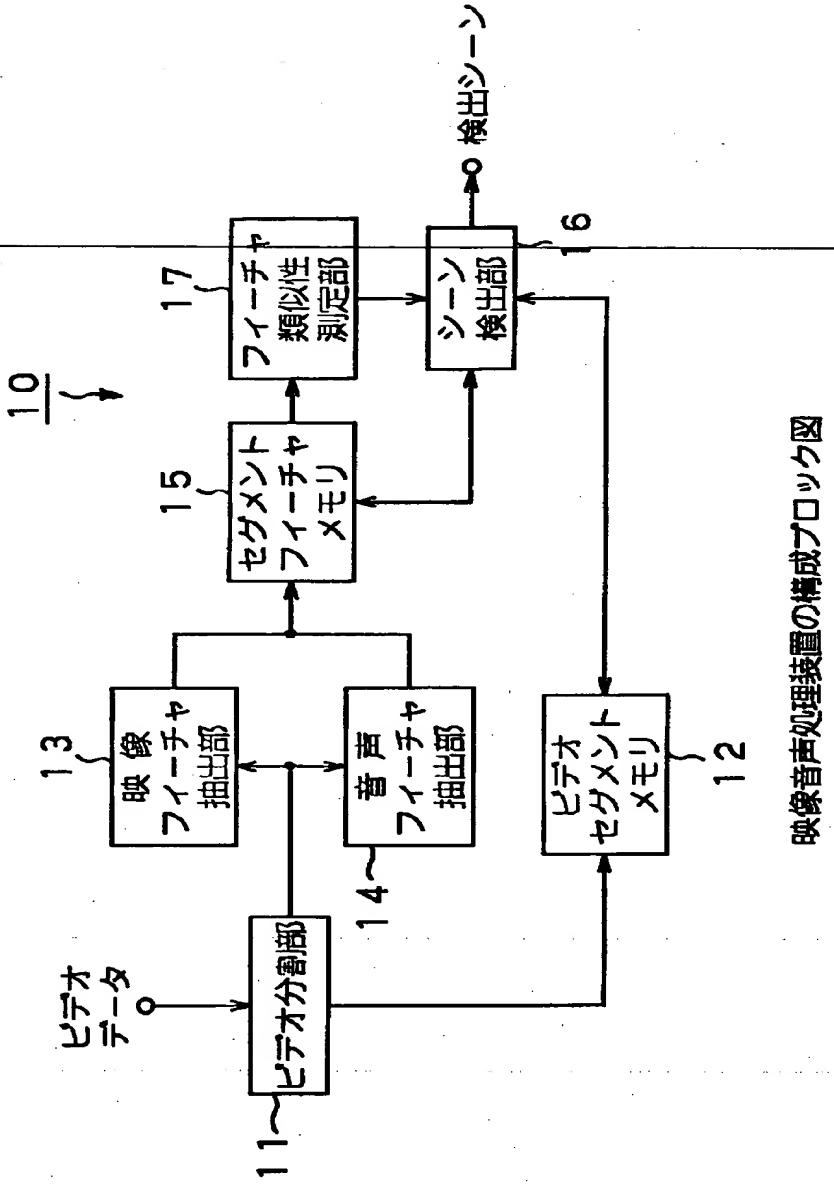
ビデオストラクチャの階層モデル

【図２】



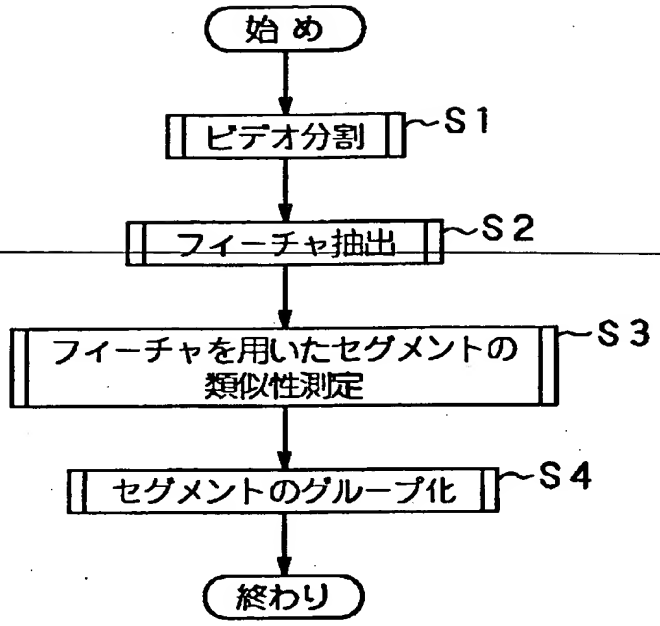
シーンの説明図

【図３】



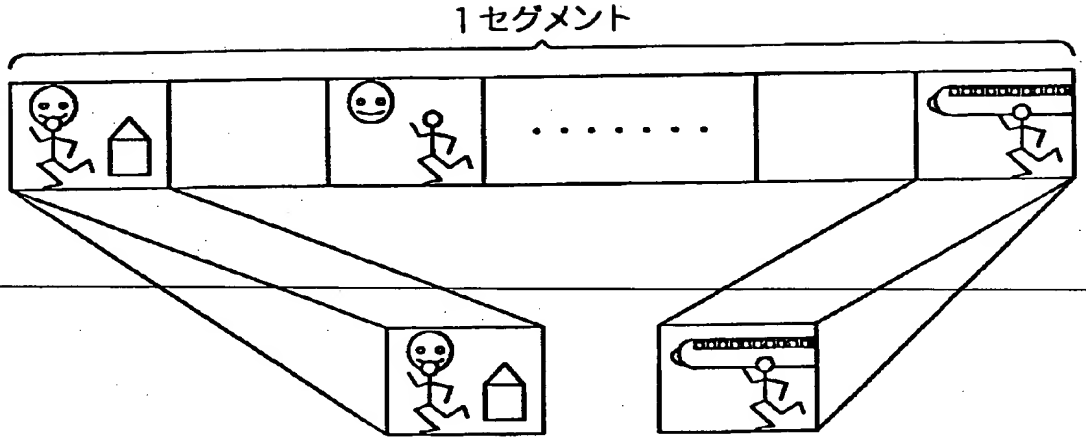
映像音声処理装置の構成ブロック図

【図４】



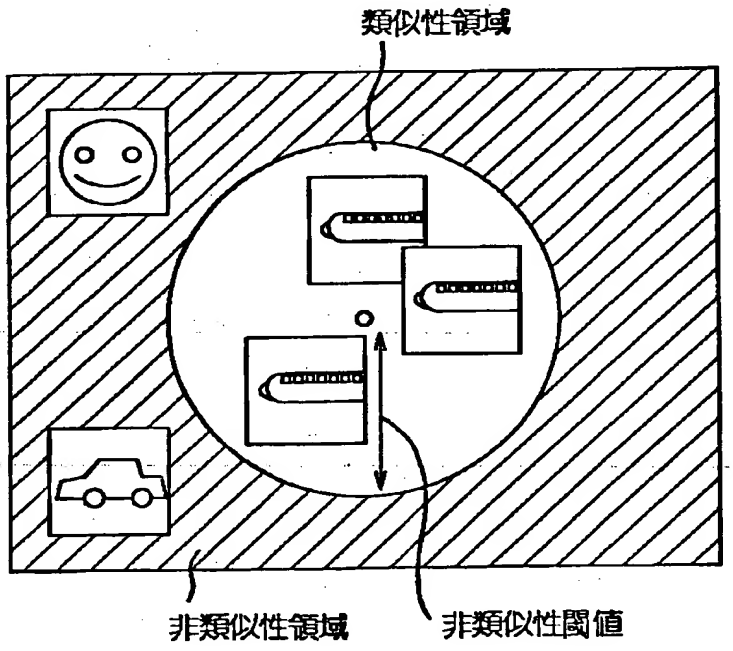
映像音声処理装置における一連の処理工程

【図５】



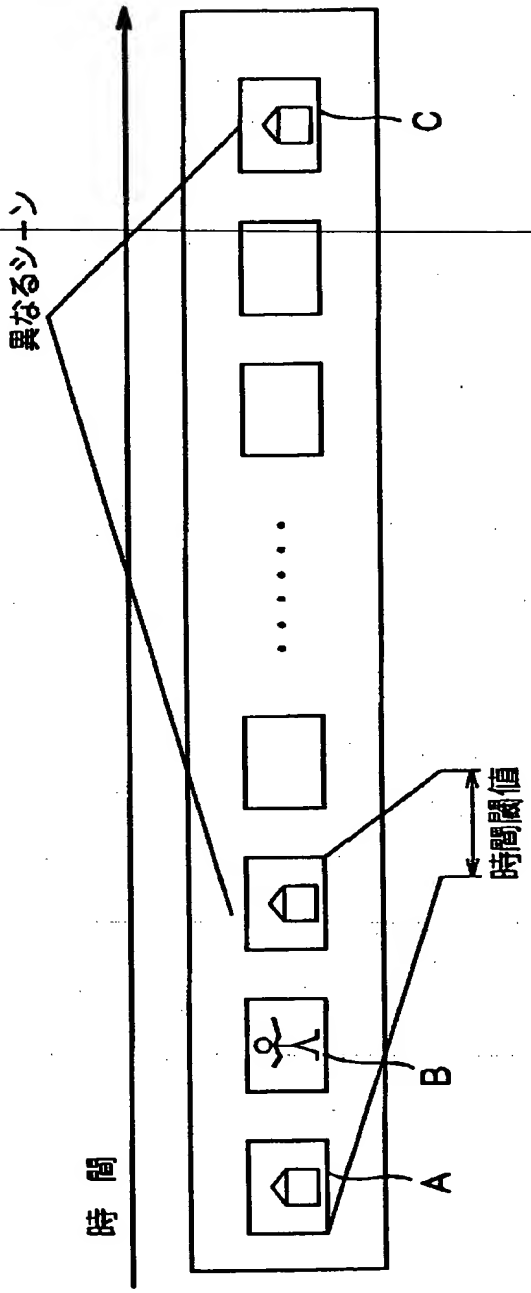
フィーチャのサンプリング方法の説明図

【図６】

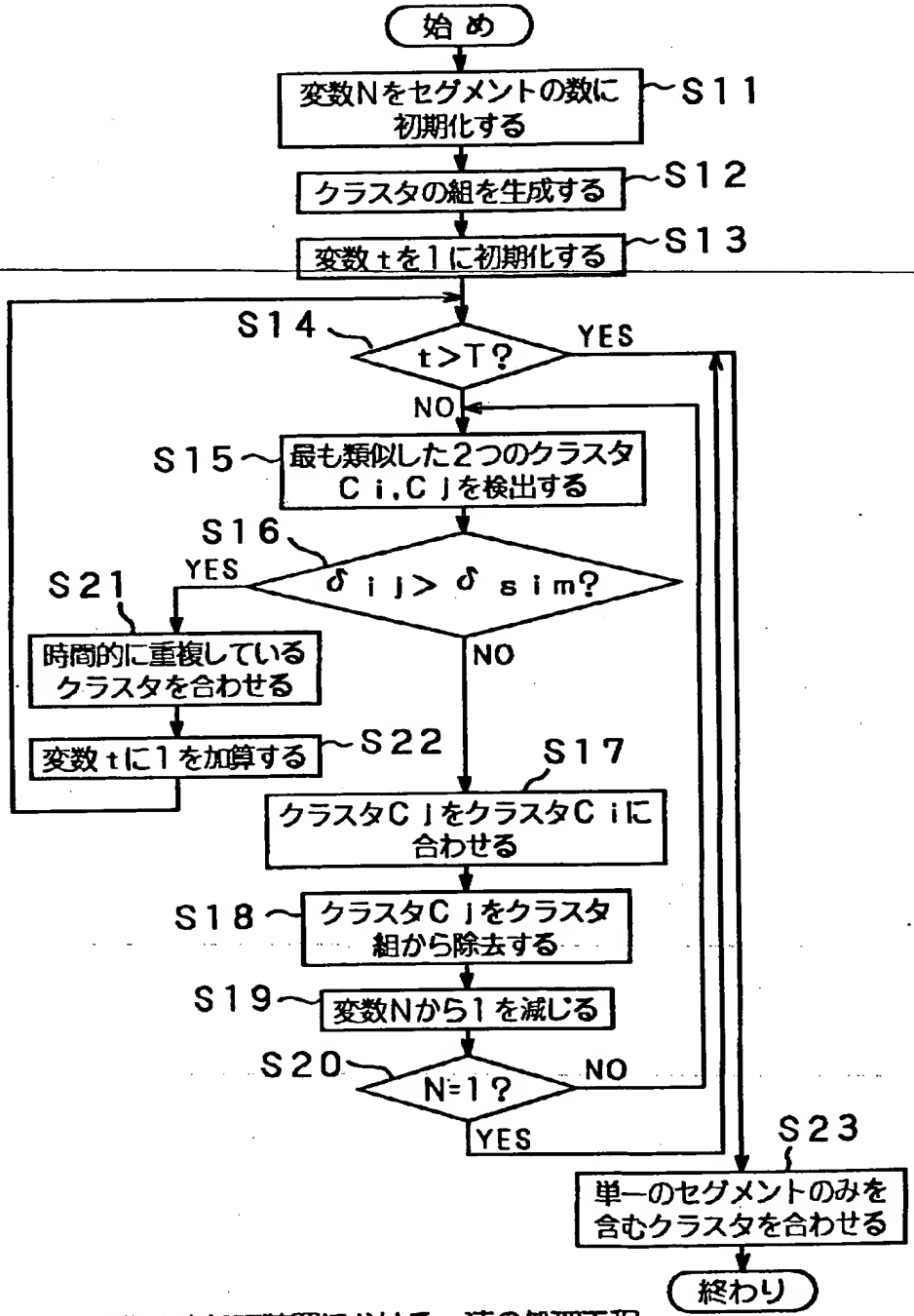


非類似性閾値の説明図

【図７】



【図８】



映像音声処理装置における一連の処理工程

【書類名】　要約書

【要約】

【課題】　種々のビデオにおける高レベルのビデオストラクチャを抽出する。

【解決手段】　映像音声処理装置１０は、入力したビデオデータのストリームから分割された映像セグメント及び／又は音声セグメントから抽出されたフィーチャと、このフィーチャを用いて、各フィーチャ毎に計算された、映像セグメント及び／又は音声セグメントの対の間の類似性を測定する測定基準とを用いて、映像セグメント及び／又は音声セグメントのうち、互いの時間的距離が所定の時間閾値以内であるとともに、互いの類似性が所定の非類似性閾値以上である２つの映像セグメント及び／又は音声セグメントを検出し、ビデオデータの内容の意味構造を反映するシーンにまとめるシーン検出部１６を備える。

【選択図】　図３

認定・付加情報

特許出願の番号	平成１１年　特許願　第０２３０６４号
受付番号	５９９０００７９３２３
書類名	特許願
担当官	第一担当上席　　００９０
作成日	平成１１年　２月１３日

＜認定情報・付加情報＞

【提出日】　平成11年 1月29日

次頁無

出願人履歴情報

識別番号　〔０００００２１８５〕

１．変更年月日　１９９０年　８月３０日
［変更理由］　新規登録
住　所　東京都品川区北品川６丁目７番３５号
氏　名　ソニー株式会社